

СЕКЦІЯ XIII. СИСТЕМНИЙ АНАЛІЗ, МОДЕЛЮВАННЯ ТА ОПТИМІЗАЦІЯ

ДОСЛІДЖЕННЯ МОЖЛИВОСТЕЙ PROMPTING ПІДХОДУ З ВИКОРИСТАННЯМ GPT 3.5-TURBO ДЛЯ ТЕКСТОВОЇ КАТЕГОРИЗАЦІЇ

Волощук Юрій Олексійович

магістрант факультету інформаційних технологій
ДВНЗ «Ужгородський національний університет», м. Ужгород, Україна

Науковий керівник: Міца Олександр Володимирович

ORCID ID: 0000-0002-6958-0870

Д-р. техн. наук, професор, завідувач кафедри інформаційних
управляючих систем та технологій

ДВНЗ «Ужгородський національний університет», м. Ужгород, Україна

Моделі машинного навчання революціонізували галузь текстової категоризації, автоматизуючи процес призначення попередньо визначених категорій або міток текстових даних [1]. Традиційні алгоритми машинного навчання, такі як наївний байес, методи опорних векторів (SVM) [2] та випадковий ліс широко використовуються для завдань текстової класифікації. Ці алгоритми ґрунтуються на інженерії ознак, де відповідна інформація витягається з тексту, часто за допомогою методів, таких як TF-IDF (Частота слів - Інверсна частота документа) або представлення "мішок слів". Хоча ці методи довели свою ефективність, вони вимагають експертизи в галузі та ручного вибору ознак, що обмежує їх адаптивність до різних текстових даних [3].

У порівнянні з цим великі мовні моделі (LLM), такі як GPT-3, GPT-3.5 та їхні наступники, виявилися потужними інструментами в обробці природної мови. Ці моделі попередньо навчені на великому обсязі текстових даних та проявили надзвичайні здатності в різних завданнях NLP, включаючи текстову категоризацію. Вони можуть автоматично вивчати контекстуальні ознаки з сирого тексту, усуваючи необхідність в явній інженерії ознак. Інтеграція LLM в завдання текстової категоризації має потенціал спростити процес та покращити точність, що робить їх захоплюючим напрямком для дослідження.

Задача категоризації є дуже актуальною і для інноваційних технологічних розробок науковців Ужгородського національного університету, таких як інтерактивна карта діалектів [4], інтерактивна карта в рамках реалізації актуальних державних та громадських напрямків [5] та платформи корпусів текстових співів церковнослов'янською мовою [6].

У цій роботі досліджено можливості prompting підходу з використанням GPT 3.5-turbo щодо його продуктивності, масштабованості та адаптивності до різних сценаріїв текстової категоризації.

Було використано набір даних із платформи, яка зв'язує дві категорії користувачів:

1. відвідувачів платформи, які мають якесь питання чи проблему і потребують допомоги кваліфікованого спеціаліста;

2. експертів, які мають кваліфікацію в різних категоріях і надають консультації.

Набір даних містить дві колонки:

1. питання користувачів англійською мовою;

2. категорія цього питання.

Попередньо було досліджено, що модель GPT 3.5-turbo виконала завдання категоризації набагато краще за davinci та curie. Проте її не вдалося протестувати з першого разу на повному наборі даних тестових даних з більш ніж 11 000 питань. Два рази програма зупинялася через перевищення часу очікування відповіді від API – один раз на прогресі 12% і один раз на 56%. У результаті оригінальний тестовий набір даних було розбито на 10 менших наборів з близько 1000 записів у кожному наборі. Потім результуючі 10 документів із передбаченими категоріям зведено в один і розраховано метрики для категоризації. Загальне успішне виконання програми категоризації для всіх документів зайняло більше двох годин.

Результати категоризації за допомогою prompting підходу з використанням GPT 3.5-turbo: Precision – 77%, Recall – 70%, F1 Score – 69%. Проаналізувавши детальніше всі випадки, коли GPT 3.5-turbo неправильно визначив категорію, виявлено, що в 3% випадків відповідь моделі виходила за межі списку наперед визначених категорій.

Висновки. Провівши дослідження, prompting підхід з використанням GPT 3.5-turbo для категоризації можна порекомендувати для наступних випадків:

- Для нових додатків, які ще не мають історичних даних для тренування моделей.

- Для компаній, які не мають ресурсів для розгортання, налаштування і підтримки моделей машинного навчання. Вартість використання LLM пропорційна активності користувачів на сайті і розтягнута в часі. У той же час, для розгортання повноцінної інфраструктури для класичних моделей машинного навчання потрібна початкова інвестиція.

- Для додатків де вхідні і вихідні дані категоризації досить малі, загальна потреба в кількості категоризацій за проміжок часу невелика і точність категоризації не є критичною для функціонування бізнесу. У таких випадках використання LLM API може бути дешевшим за повноцінну ML інфраструктуру.

- Для додатків, де немає потреби вибирати із наперед визначеного списку категорій або для випадків коли цей список потрібно дуже часто оновлювати і перетреноувати. Для таких додатків LLP є більш гнучким рішенням, що краще адаптується до змін вхідних даних.

Список використаних джерел:

1. Lupei, M., Mitsa, O., Sharkan, V., Vargha, S., & Gorbachuk, V. (2022, April). The identification of mass media by text based on the analysis of vocabulary peculiarities using support vector machines. In 2022 International Conference on Smart Information Systems and Technologies (SIST) (pp. 1-6). IEEE. doi: 10.1109/SIST54437.2022.9945774.
2. Lupei, M., Mitsa, O., Sharkan, V., Vargha, S., Lupei, N. (2023). Analyzing Ukrainian Media Texts by Means of Support Vector Machines: Aspects of Language and Copyright. In: Hu, Z., Dychka, I., He, M. (eds) Advances in Computer Science for Engineering and Education VI. ICCSEE 2023. Lecture Notes on Data Engineering and Communications Technologies, vol 181. Springer, Cham. doi: 10.1007/978-3-031-36118-0_16.
3. Lupei, M., Mitsa, A., Povkhan, I., & Sharkan, V. (2020, August). Determining the eligibility of

- candidates for a vacancy using artificial neural networks. In 2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP) (pp. 18-22). IEEE. doi: 10.1109/DSMP47368.2020.9204020.
4. Mitsa, O., Sharkan, V., Maksymchuk, V., Varha, S., & Shkurko, H. (2023). Ethnocultural, Educational and Scientific Potential of the Interactive Dialects Map. In 2023 IEEE International Conference on Smart Information Systems and Technologies (SIST) (pp. 226-231). Astana, Kazakhstan. doi: 10.1109/SIST58284.2023.10223544.
 5. Lupei, M., Shlahta, M., Mitsa, O., Horoshko, Y., Tsybko, H., & Gorbachuk, V. (2022, September). Development of an Interactive Map Within the Implementation of Actual State and Public Directions. In 2022 12th International Conference on Advanced Computer Information Technologies (ACIT) (pp. 384-387). IEEE. doi: 10.1109/ACIT54803.2022.9913191.
 6. Popovych, N., Lutskiv, A., Mitsa, O., Lyntvar, O., & Ivanova, A. (2023). Ukrainian Redaction of Church Slavonic (URCS): Needs for Digitalization and Text Corpora Platform Generation. Part I.